DOCUMENT RESUME

ED 464 944 TM 033 875

AUTHOR King, Jason E.
TITLE Bootstrapping Confidence Intervals for Robust Measures of Association.
PUB DATE 2002-02-00
NOTE 41p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, February 14-16, 2002). Based on the author's Doctoral Dissertation, Texas A&M University, 2000.
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Correlation; Monte Carlo Methods; *Robustness (Statistics); Simulation
IDENTIFIERS *Bootstrap Methods; *Confidence Intervals (Statistics); Measures of Association

ABSTRACT

A Monte Carlo simulation study was conducted to determine the bootstrap correction formula yielding the most accurate confidence intervals for robust measures of association. Confidence intervals were generated via the percentile, adjusted, BC, and BC(a) bootstrap procedures and applied to the Winsorized, percentage bend, and Pearson correlation coefficients. Type I error, bias, efficiency, and interval length were compared across correlational and bootstrap methods. Results reveal the superior resiliency of the robust measures over the Pearson r, though neither robust correlation outperformed the other. Unexpectedly, the four bootstrap techniques achieved roughly equivalent outcomes. Based on these results, it appears that the more complex bootstrapping procedures may not be worth the additional computational expenditures. (Contains 6 tables, 2 figures, and 76 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 464 944

Running Head: BOOTSTRAPPING CONFIDENCE INTERVALS

J. King

1

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Bootstrapping Confidence Intervals for Robust Measures of Association

Jason E. King

Baylor College of Medicine

TM033875

Paper presented at the annual meeting of the Southwest Educational Research Association, February, 2002. This paper was based on a doctoral dissertation.

Correspondence concerning this article should be addressed to Jason King, 1709 Dryden Suite 534, Medical Towers, Houston, TX. 77030. E-mail: Jasonk@bcm.tmc.edu

BEST COPY AVAILABLE

Abstract

A Monte Carlo simulation study was conducted to determine the bootstrap correction formula yielding the most accurate confidence intervals for robust measures of association. Confidence intervals were generated via the percentile, adjusted, BC, and BC_a bootstrap procedures and applied to the Winsorized, percentage bend, and Pearson correlation coefficients. Type I error, bias, efficiency, and interval length were compared across correlational and bootstrap methods. Results revealed the superior resiliency of the robust measures over the Pearson r, though neither robust correlation outperformed the other. Unexpectedly, the four bootstrap techniques achieved roughly equivalent outcomes. Based on the these results, it appears that the more complex bootstrapping procedures may not be worth the additional computational expenditures.

Bootstrapping Confidence Intervals for Robust Measures of Association

The purpose of this study was to compare bootstrapping approaches to estimating confidence intervals for various correlation coefficients. The paper initially describes the concept of robustness, robust and non-robust measures of association, and bootstrapping procedures. Next, the simulation method is detailed, including a discussion of indices used to compare the various measures and procedures. Finally, results conclusions, and recommendations are offered.

The Concept of Robustness

Classical statistical procedures are often inadequate due to their sensitivity to departures from distributional assumptions. The extent to which an estimator is able to withstand such deviations has been dubbed robustness (Box, 1953). The term robustness is here used in the narrow sense as applied only to distributional assumptions, though other standard assumptions could be invoked. Although conceptually distinct, distributional robustness and outlier resistance are essentially synonymous notions (Huber, 1981).

Robustness, or resistance as it was initially termed, was generally understood from the inception of the major statistical advances that occurred during the 19th and early 20th centuries but was not seriously examined until the 1950s (Staudte & Sheather, 1990; Stigler, 1973). While some earlier theorists assessed the consequences of distributional nonnormality for hypothesis testing (viz., the robustness of validity), few explored the stability of power or of the length of confidence intervals (viz., the robustness of performance), though the latter "usually [brings] for free a satisfactory robustness of validity (but not vice versa)" (Huber, 1972, p. 1046). Tukey counseled against the then-prevailing habit of "sweeping the dirt under the rug" (1960, p. 450) and ignoring comparisons of the relative efficiency of various point estimates. He admonished, "nearly imperceptible non-normalities may make conventional relative efficiencies of estimates of scale and location entirely useless" (p. 474).

Partly as a result of such admonitions, a number of robust analogs to traditional estimators, population parameters, and hypothesis-testing methods have been developed during the past 40 years. Robust procedures typically retain the statistical interpretations associated with classical procedures but are more resistant to data nonnormality.

Nevertheless, applied researchers have been slow to adopt the newer methods. The hesitancy is due in part to a lack of knowledge. Over two decades ago, Bradley (1978) bemoaned the general poor treatment of assumption violation in elementary statistics textbooks noting that "reassuring complacency of tone, depreciating the consequences of assumption-violation, or using overly exuberant language to exult over claimed robustness seems to be endemic in statements about robustness" (p. 145). It is no wonder that "most [modern] psychologists believe that all practical problems associated with statistical methods were solved by the year 1955" (Wilcox, 1998b, p. 60). In spite of this widespread misconception, several have recently asserted the need for newer methods (e.g., Wilcox, 1996, 1998a; on the debate in general, see the May 1998 issue of the British Journal of Mathematical & Statistical Psychology, as well as Hampel's [1998] depiction of the current state of affairs).

The Pearson Product-Moment Correlation

One common misconception involves the robustness of the product-moment correlation coefficient. The classical functional for measuring linear association between two continuously-scaled variables is defined as

$$\rho_{XY} = \frac{\Sigma(x - \mu_x)(y - \mu_y)}{N\sigma_x\sigma_y} , \qquad (1)$$

where μ_x and μ_y are the population means of the population variables $\underline{x}$ and $\underline{y}$, respectively, and σ_x and σ_y are the population standard deviations of $\underline{x}$ and $\underline{y}$, respectively. The maximum likelihood sample estimator of rho, $\hat{\rho}$, is estimated by $\underline{r}$:

$$\hat{\rho}_{XY} = r_{XY} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{ns_Xs_Y} , \qquad (2)$$

where $\bar{X}$, $\bar{Y}$, $\underline{s}_X$, and $\underline{s}_Y$ are the sample means and standard deviations of the sample variables $\underline{X}$ and $\underline{Y}$, respectively.

Sir Ronald A. Fisher defined the normal theory sampling distribution of $\underline{r}$ in samples of size $\underline{n}$ (1915, p. 508) to be

$$f(r_{XY}) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1 - \rho^2}} e^{-\frac{1}{1-\rho^2}\left[\frac{(X-\mu_X)^2}{2\sigma_X^2} - \frac{2\rho(X-\mu_X)(Y-\mu_Y)}{2\sigma_X\sigma_Y} + \frac{(Y-\mu_Y)^2}{2\sigma_Y^2}\right]} dxdy \quad . (3)$$

With "large samples and moderate or small correlations" (Fisher, 1958, p. 192) the sample correlation coefficient is distributed normally around ρ with variance

$$\text{var}(r) = \frac{(1 - \rho^2)^2}{n - 1} \quad . \tag{4}$$

The standard $(1 - \alpha)100\%$ confidence interval limits for ρ, again assuming sampling distribution normality, are calculated by

$$\text{Lower} = r - z_{1-\alpha/2} * \sqrt{\text{var}(r)} \tag{5}$$

and

$$\text{Upper} = r - z_{\alpha/2} * \sqrt{\text{var}(r)} \quad , \tag{6}$$

where α is the inclusion probability. Here, $z_{1-\alpha/2}$ and $z_{\alpha/2}$ are quantiles from the standard normal distribution.

Fisher's *z* Transformation

Because the sampling distribution of $\underline{r}$ is complicated when $\rho_{XY} \neq 0$, a transformation of $\underline{r}$ was proposed by Fisher (1915, 1921):

$$\hat{\rho}_z = r_z = z' = \tanh^{-1} r \, , \qquad \rho_z = \zeta = \tanh^{-1} \rho \quad , \tag{7}$$

or equivalently

$$z' = .5 \frac{\ln(1 + r)}{\ln(1 - r)} \, , \qquad \zeta = .5 \frac{\ln(1 + \rho)}{\ln(1 - \rho)} \quad . \tag{8}$$

This inverse hyperbolic tangent transformation compensates for the nonnormality of the sampling distribution of $\underline{r}$ and approaches a normal distribution.

A $(1 - \alpha)100\%$ confidence interval for $\underline{z}'$ is formed as

$$\text{Lower} = \left(z' - \frac{.5\hat{\rho}}{n - 1} \right) - z_{\alpha} (n - 3)^{-.5} \tag{9}$$

and

$$\text{Upper} = \left(z' - \frac{.5\hat{\rho}}{n-1}\right) + z_{\alpha}(n-3)^{-.5} \quad , \qquad (10)$$

where z_{α} is the α quantile of the standard normal distribution and α the inclusion probability. The intervals can be reverse-transformed to the metric of r for ease of interpretation.

Studies on the Robustness of *r*

An extensive literature review of the robustness of r was conducted, but will not be presented here due to space limitations (for details, see King, 2000). To summarize, approximately two-thirds of the reviewed studies expressed reservations for the use of r under some nonnormal conditions. That percentage was even higher when only more recent studies were considered. Due to the availability of more powerful simulation capabilities beginning around 1960, studies conducted after that date should be weighed more heavily.

Under independence (i.e., $\rho = 0$) most reported the sampling distribution of r to be robust to bivariate nonnormality (Duncan & Layard, 1973; Dunlap, 1931; Gayen, 1951; Havlicek & Peterson, 1976, 1977; Nair, 1941; Norris & Hjelm, 1961; Pearson, 1931, 1932), unless mixed or contaminated distributions were under review (Devlin, Gnanadesikan, & Kettenring, 1975; Duncan & Layard, 1973; Edgell & Noon, 1984; Kowalski, 1972; but cf. Pearson, 1929; Srivastava & Awan, 1984). However, the majority of studies employed distributions that diverged minimally from the bivariate normal condition, especially the earlier studies (e.g., Pearson, 1931, reported skewness = .99 and kurtosis = 3.83 for his most excessive condition!). For more extreme departures like severe skewness (Baker, 1930), the L distribution (Blair & Lawson, 1982), or the Cauchy distribution (Edgell & Noon, 1984), r was <u>not</u> found to be robust.

When ρ does not equal zero and the bivariate surface is nonnormal, r is likely biased, sometimes inordinately so. Of all the reviewed studies which included a dependence condition, only two (Pearson, 1929; Zeller & Levin, 1974) found r to be completely robust; the majority expressed reservations for at least some situations (Cheriyan, 1945; Chesire et al., 1932; Devlin et al., 1975; Duncan & Layard, 1973; Gayen, 1951; Haldane, 1949; Hey, 1938; Kowalski, 1972; Norris & Hjelm, 1961; Rider's [1932] results, though not his comments).

Numerous researchers hold that all uses of r are completely robust to distributional assumptions. This misunderstanding may have arisen due to the selective literature reviews conducted in some prominent studies. For example, Edgell and Noon (1984)

failed to survey robustness studies conducted by Baker (1930) and Blair and Lawson (1982). Both established the inadequacy of *r* under certain nonnormal conditions, including the case of $\rho = 0$. In addition, certain literature reviews neglected to cite any studies demonstrating the non-robustness of *r* (e.g., Zeller & Levine, 1974) or failed to accurately represent such findings in summarizing their literature reviews (e.g., Havlicek & Peterson, 1976, p. 1321).

Further, when quantitative measures of resistance are applied to ρ, additional problems surface. For example, the influence function and breakdown point of ρ suggest that even a single pair of outlying scores can render the parameter virtually meaningless for quantifying the bivariate relationship underlying the majority of data points (Devlin, et al., 1975; Wilcox, 1993).

Robust Measures of Correlation

The literature on robust correlations is scattered and scarce, especially regarding comparative simulation studies. Though the quadrant correlation has been available for a century (Blomqvist, 1950; Sheppard, 1899), the majority of resilient measures of association have been introduced only recently. Of these, two appear particularly promising. Along with possessing properties that curb the influence of distributional anomalies, both the Winsorized correlation (Devlin et al., 1975; Gnanadesikan & Kettenring, 1972; Wilcox, 1993) and the percentage bend correlation (Wilcox, 1994) yield interpretations analogous to the popular Pearson *r*. Yet few have explored these newer correlation coefficients, notably with respect to generating accurate confidence intervals. Likewise, little is known about the latent sampling distribution of each index.

The Winsorized Correlation

Winsorization involves ordering the scores in a distribution and then deleting the γ smallest values and setting them equal to $X_{(\gamma + 1)}$, and deleting the γ largest values and setting them equal to $X_{(n - \gamma)}$. In other words, potential outliers are removed from each tail of the distribution and replaced with the most extreme score remaining in that tail. Delvin et al. (1975) present the equations and proofs for calculating a sample Winsorized correlation coefficient (r_w) based on the Winsorized sample mean and the Winsorized sample variance. This statistic can be conceptualized as the application of the ordinary Pearson formula to two Winsorized score distributions.

Wilcox (1993) evaluated Type I error probabilities and power calculations for the ordinary r versus r_w via computer simulations. The Winsorized correlation bettered r in terms of error rates across a range of conditions. For the bivariate normal case r evidenced superior power; however, r_w demonstrated greater power when nonnormal conditions were explored. Wilcox could not determine an optimal method for generating accurate confidence intervals for the dependence condition, including application of the bootstrap (though no results were offered), but some success was had in transforming r_w to a regression coefficient.

The Percentage Bend Correlation

Wilcox (1994) suggested a percentage bend correlation as a robust measure of association. This correlation is based on the percentage bend measure of location and the percentage bend midvariance. Computational details and proofs are discussed in Wilcox (see also King, 2000). Outlying scores are defined via a constant labeled β. As with the Pearson and Winsorized correlations, the percentage bend correlation will equal zero (or come very close) under independence and fall between -1 and +1. Many alternative robust correlations do not meet these criteria.

Few studies have investigated this new correlation coefficient. Wilcox (1994) defined the measure and provided a test for independence. He also compared tests of statistical significance for r, r_w, and r_{pb}. In comparison with r, tests of r_{pb} more closely mirrored expected Type I error rates for samples of size 10 and 20 with $\alpha_{nominal} = .05$, except under bivariate normality. In terms of power evaluations, tests of r_{pb} compared favorably with tests of r and generally outperformed tests of r_w under nonnormal conditions, so long as $\beta = .1$ was used to compute r_{pb}. However, the Pearson correlation edged out the others under distributional normality.

Bootstrapping Confidence Intervals

For statistics with no known sampling distribution, Efron's (1979, 1982) bootstrap has proven to be effective in a variety of contexts. The conjecture is that the sampling distribution of a statistic can be approximated by the distribution of a large number of resampled estimates of the statistic obtained from a single sample of observations. The distribution of resampled estimates forms an empirically-derived sampling distribution from which confidence intervals or other indices may be estimated.

Efron settled on the name from the mythical Baron von Munchausen's unique manner of escape from a deep lake: He pulled himself up by his bootstraps (Efron & Tibshirani, 1993). Tukey's suggested nomenclature was equally descriptive. According to Efron (1979, p. 25), Tukey favored the term shotgun because the method "can blow the head off any problem if the statistician can stand the resulting mess."

The bootstrap sampling distribution may be employed either for inferential purposes (e.g., testing hypotheses about parameters) or for description (e.g., estimating a likely range for some parameter at a given confidence level or result stability or replicability) (Thompson, 1993). The technique is especially valuable when confronted with statistics having no known sampling distribution.

Due to its efficacy, bootstrapping has become fashionable in many fields. Wilcox (1997a, p. 45) stated that over 1,000 journal articles on bootstrapping have already seen publication! Further, "an almost bewildering array" of variants of bootstrap confidence intervals have been advanced (Hall, 1988, p. 927). These vary in the accuracy with which the bootstrap-generated interval spans the true interval. Accuracy is also contingent on the statistic under examination: No single bootstrapping routine is optimal across all statistical techniques.

The present investigation considered four of the more popular, well-studied procedures: the percentile bootstrap, the adjusted bootstrap, the bias-corrected bootstrap, and the bias-corrected and accelerated bootstrap.[1] Though some of the newer, more complex methods appear promising, it seems that most remain largely in the developmental stage or require unreasonable execution time. Additional research into the theoretical properties and practical usage of these newer approaches is still needed.

Description of the Bootstrap

Let $\underline{x}$ be a population of observations and $\underline{X}$ a random sample of $\underline{n}$ observations drawn from population $\underline{x}$. A parameter summarizing the population will be denoted as θ, and the sample estimate as $\hat{\theta}$. The symbol $F(\hat{\theta})$ will represent the theoretical sampling distribution for the population.

In Monte Carlo bootstrapping experiments, each of $\underline{m}$ simulated data sets (samples) are resampled $\underline{B}$ times, with each resample being of size $\underline{n}$ (Lunneborg, 2000). Resampling involves repeatedly drawing observations from a sample with replacement until $\underline{n}$ is reached. Next, a bootstrap estimate of the parameter is acquired for each resample. The bootstrap estimate is

represented by $\hat{\theta}^*$. The process is then repeated B times yielding a bootstrap sampling distribution of estimated parameters (denoted $\hat{F}(\hat{\theta}^*)$) from which confidence intervals and other information can be derived. In theory the bootstrap sampling distribution should mimic the true sampling distribution of the statistic.[2]

Types of Bootstrapped Confidence Intervals

The percentile bootstrap (Efron, 1979) ranks as the most popular bootstrap procedure (according to Hall's, 1988, informal inquiries), although "the great majority of non-technical statistical work...does not make it clear which...[technique] is employed" (Hall, p. 927). This "arch" bootstrap (Sievers, 1996, p. 381) entails first calculating $\hat{\theta}^*$ for each resample. The distribution of conditionally independent statistics, $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$, are then ordered by value. An α-level confidence interval includes all the values of $\hat{\theta}^*$ between the $\alpha/2$ and the $1 - \alpha/2$ percentiles of the (ordered) bootstrap sampling distribution, $\hat{F}(\hat{\theta}^*)$.

Another option is to widen the traditional bootstrap confidence band by the factor $[(n+2)/(n-1)]^{1/2}$ (Efron, 1982). Both endpoints of the confidence interval are adjusted an equal amount. Following Strube (1988), this approach will be here referred to as the adjusted bootstrap.

Third, Efron (1981, 1982, 1985) offered the bias-corrected (BC) bootstrap to correct for problems with the percentile bootstrap. The BC method allows for the possibility that the distribution of $\hat{\theta}^* - \hat{\theta}$ is not centered on zero but is distributed around a constant $z_0\hat{\sigma}_{\hat{\theta}}$, where $\hat{\sigma}_{\hat{\theta}}$ is the standard error of $\hat{\theta}$. In other words, the procedure rests on the assumption that there exists some monotonic transformation of $\hat{F}(\hat{\theta})$ whose differences are distributed in some known way, usually as a normal variate (hence the constant being labeled z_0). The constant must next be estimated and applied to the bootstrap sampling distribution. Because the method is transformation invariant, it is not required that the specific transformation function be known, only that one exists (Mooney & Duval, 1993).

Efron (1987) proposed the bias corrected and accelerated (BC_a) bootstrap to stabilize the variance of the bootstrap distribution (i.e., remove the tendency for the variance to accelerate across values in the sampling distribution), in addition to incorporating the bias correction described above.

Extensive calculations are involved with this procedure (but see Lunneborg [2000, pp. 164-165] for a readable description).

In spite of the involved computational demands, the BC_a holds much promise as a bootstrap correction procedure for several reasons. In terms of coverage error, Efron (1987) and Hall (1988) proved that the BC_a method yields second-order correctness under a wider class of problems than the BC bootstrap. Second, the BC_a bootstrap is transformation invariant. Moreover, the specific transformation may remain unknown while employing the BC_a.

However, there are potential drawbacks to the method. First, it is assumed that a transformation exists that yields a distribution of differences with a known shape (Mooney & Duval, 1993). Second, Hall (1988) lamented the fact that standard normal tables must be consulted in applying the method. But with the advent of more complex statistical computer programs, this deterrence is soon becoming obsolete. Third, the acceleration constant proves troublesome to calculate in many cases (DiCiccio & Romano, 1989). Finally, simulations conducted by Hall suggest that the BC_a bootstrap may produce abnormally short intervals with small samples when wide nominal coverage is of interest.

Research on Bootstrapping *r*

A thorough literature review was conducted of studies that applied bootstrapping to the Pearson correlation. These results will be briefly summarized (for details, see King, 2000). Although results were somewhat mixed, bootstrapping appears to improve upon the normal theory intervals both in terms of probability coverage and interval accuracy for most mixed normal and nonnormal distributional conditions (Efron, 1988; Hall, Martin, & Schucany, 1989; Lunneborg, 1985; Rasmussen, 1988, 1989; Sievers, 1996; Strube, 1988). But under bivariate normality the standard intervals may be preferable (Rasmussen, 1987). While some studies reported problems with statistical significance levels when stringent nominal alphas were employed, others found the opposite effect (cf. the 1987 and 1989 experiments both conducted by Rasmussen!). Methodological variations may be the source of such discrepancies.

In comparing the various bootstrapping procedures, it seems that the adjusted intervals (i.e., adjusted, BC, and iterated bootstrap[3]) formed more accurate confidence intervals and probability levels than did the percentile and percentile-*t*[3] methods (e.g., Hall et al., 1989; Strube, 1988). Results for the latter two techniques differed in that the percentile tended to undercover while the percentile-*t* tended to overcover (Hall et al., 1989; Rasmussen, 1987; Sievers, 1996). The ordinary

percentile-t was found to effect relatively large standard errors (Hall et al., 1989; Sievers, 1996), though a modified version performed well. Finally, a study by Wilcox (1991) in which no bootstrap method achieved accurate intervals under variable dependence seems to be somewhat of an anomaly.

While it seems that at least some problems with testing and estimating ρ can be alleviated by applying the bootstrap, others cannot. Even if a bootstrap method proves valuable in obtaining proper confidence intervals for ρ, one may not wish to use a correlation coefficient that is so affected by outliers and related distributional departures from normality.

Research on Bootstrapping Robust Correlations

Little research has been done on the application of the bootstrap to robust correlations. Wilcox (1997b) proposed a test for the independence of $p > 2$ random variables using the percentage bend correlation. Although this multivariate application is not germane to our study, he also examined the computation of confidence intervals for ρ_{pb} in the bivariate case. A percentile bootstrap method was employed in which $B = 399$ resamples were simulated from theoretical distributions demonstrating varying levels of skewness and kurtosis (in some cases reaching very extreme conditions). One thousand samples were acquired per condition. Population correlations of $\rho = 0$, .3, .6, and .9 were examined. Bootstrapping was also applied to ρ for comparative purposes.

Because the percentage bend correlation does not exactly equal ρ under dependence, approximate populations were generated by drawing 100,000 pairs of samples and then calculating ρ_{pb} for each. The bootstrap confidence intervals for ρ_{pb} more closely spanned the desired interval for practically all conditions and sample sizes than did those computed for ρ. The latter diverged considerably under extreme nonnormality. For example, with $n = 50$, $\rho = .9$, $\alpha = .05$, and high kurtosis, the confidence interval for ρ_{pb} yielded an interval with probability coverage of $1 - \alpha = .966$, while the interval for ρ only covered at .253. Though results were not provided, Wilcox assessed conditions in which one or both marginal distributions were skewed, sometimes in opposite directions. These outcomes were (reportedly) promising as well.

Wilcox (2001) applied the percentile bootstrap, using $B = 599$, to two measures of association: ρ and Spearman's rank correlation (ρ_s). He also examined a modified percentile method in conjunction with ρ, and a nested bootstrap applied to all

three correlation coefficients. As regards Type I error, Wilcox found that the nested bootstrap with ρ produced acceptable error probabilities, and either bootstrap method worked well with the robust correlations. However, none of the methods resulted in satisfactory confidence intervals.

Importance and Purpose of the Study

Given the current emphasis in many fields on reporting effect size measures and confidence intervals for point estimates (Thompson, 2001; Vacha-Haase, T., Nilsson, J.E., Reetz, D.R., Lance, T.S., & Thompson, B., 2000; Wilkinson & APA Task Force on Statistical Inference, 1999), estimation is clearly an important issue to consider in addition to inferential testing. But at the current level of research, it is unknown which bootstrap technique should be exercised in constructing confidence intervals for robust measures of linear relationship.

Although the Winsorized and percentage bend correlations have been compared with each other and with the ordinary correlation in terms of committing Type I errors (Wilcox, 1993, 1994, 1997b), only Wilcox (1997b, 2001) has examined the accuracy of bootstrapped confidence intervals for robust measures. Clearly, more research is needed in this area. So the primary purpose of the present study was to compare various methods of bootstrapping confidence intervals for each of the above-mentioned robust correlations and ρ. For completeness, the Fisher-transformed correlation will be included, although it frequently fails to produce even asymptotically correct results (Duncan & Layard, 1973).

Research Questions

1. How do confidence intervals for the robust correlations ρ_w and ρ_{pb} compare with those for ρ and its transform in terms of accuracy and stability?

2. Which bootstrap method provides the most accurate and stable confidence intervals for ρ_w? for ρ_{pb}?

3. With small samples which bootstrap method provides the most accurate and stable confidence intervals for ρ_w? for ρ_{pb}?

4. Under extreme distributional conditions which bootstrap method provides the most accurate and stable confidence intervals for ρ_w? for ρ_{pb}?

Method

General Simulation Procedure

Due to its computational efficiency in bootstrapping (Wilcox, 1997b), all computer simulations were conducted using the S-Plus statistical package running on a 600 MHz Pentium III computer with 256 Mb RAM. The procedure for the simulations was as follows:

1. Randomly generate N = 1,000,000 observations from a population with known characteristics (i.e., constrained through the simulation procedure to have certain parametric properties). This is the derived population. The step is necessary because the Winsorized and percentage bend correlations will not usually exactly equal ρ when dependence exists, and a simulated population allows comparisons between ρ and the robust correlations.
2. Calculate the parameters ρ, ρ_z, ρ_w, and ρ_{pb} for the population.
3. Randomly select without replacement a sample of size n from the population.
4. Calculate the statistics r, r_z, r_w, and r_{pb} for the sample.
5. Randomly select with replacement a resample of size n from the sample. This is one bootstrap sample.
6. Calculate the statistics r^*, r_z^*, r_w^*, and r_{pb}^* for the resample, where the asterisk denotes a bootstrap estimate.
7. Repeat Steps 5 and 6 a total of B = 500 times forming 500 bootstrap samples.
8. Calculate 95% confidence intervals for ρ, ρ_z, ρ_w, and ρ_{pb} using each of the four bootstrap procedures.
9. Repeat Steps 3 through 8 a total of m = 100 times forming 100 samples.
10. Repeat Steps 1 through 9 for each condition of interest (described below).

While one would normally wish to secure at least 1,000 bootstrap samples (for a fuller discussion, see King, 2000), the extensive computations required for the present study (i.e., estimation of confidence intervals via four bootstrap methods for each of four correlation coefficients per sample) precluded this as a possibility. With B set to 3,000, for example, results from only about 25 samples were acquired in eight hours time. With 100 samples needed per condition and around 150 total conditions to be evaluated, employing such a large number of resamples became impractical.

Instead, it was decided that 500 resamples would yield relatively accurate confidence intervals, at least for comparative purposes. Each bootstrap method should "suffer" equally due to the small value of B, assuming that the accuracy of each is dependent on sample size to the same extent. Similarly, it would have been desirable to generate 1,000 samples, but m = 100 was settled upon due to computational time restraints.

Population Characteristics

Real data often demonstrate excessive distributional nonnormality (Bradley, 1977; Micceri, 1989; Rasmussen, 1986; Stigler, 1973; Wilcox, 1990). In fact, after reviewing 440 empirical studies, Micceri (1989) encountered only 15.2% of the score distributions with both tails having weights at or about those for the Gaussian distribution. Various distributional abnormalities can moderate the accuracy of a bootstrap procedure for a given statistic (Hall, 1988; Wilcox, 1997b). Thus for the present study it was decided to vary distributional shape, strength of correlation, and sample size in evaluating bootstrap approaches. Contaminated and mixed distributions were also investigated.

Following Wilcox (1997b) the g and h distribution (Hoaglin, 1985) was adopted to alter the skewness and kurtosis of each population. Hoaglin's distributions allow one to construct marginals according to four general shapes: normal, symmetric with a heavy tail, asymmetric with a light tail, and asymmetric with a heavy tail. Increasing g skews X, and h influences kurtosis. When both g and h are set to zero, X has a standard normal distribution.

The method consists of initially generating observations from a standard normal distribution, Z, for each variable. X is then set to

$$X = \left(\frac{e^{gZ} - 1}{g}\right) e^{hZ^2/2} \tag{82}$$

for g > 0; otherwise

$$X = Ze^{hZ^2/2} \tag{83}$$

to prevent division by zero.

Table 1 lists summary statistics using several illustrative populations each having 1,000,000 observations for the values of

g and h that were employed in the present study. A wide range of marginal distributional shapes were investigated.

Several sample sizes were examined: n = 20, 50, 100, and 250. The larger ns represent sample sizes greater than are typically available to most social scientists (cf. Thompson, 1999; Thompson & Snyder, 1997), and samples smaller than n = 20 should probably not be employed with robust correlations.

Strength of linear relationship was also varied. Population rhos took on values of 0, .4, or 8. However, these values will only be met for the Pearson correlation because the Winsorized and percentage bend correlations will not necessarily equal ρ under dependence.

In addition, the effect of mixed and contaminated distributions was investigated. Three mixed distributions were created for each of two nonnormal distributional conditions. Various population correlations (i.e., $\rho \approx 0, .3, .5, .7$) were constructed under two distributional conditions (slight kurtosis: marginal distributions set to g = 0, h = .1; extreme nonnormality: marginals set to g = .5, h = .3). The classic case of contamination is formed by combining a normal marginal distribution (X) and its square (X^2). This is a stringent test of any index because the distributions are completely dependent but also uncorrelated (Edgell & Noon, 1984). This condition was evaluated with n set to 100.

Comparative Criteria

Type I error. Type I error rate, interval ratio, bias, and standard error estimates were used for comparing the performance of the bootstrap methods. Observed Type I error rates were recorded using the equation $p_{obs} = m^{-1}\#\left(\hat{\theta}_{lower} \leq \theta \leq \hat{\theta}_{upper}\right)$, where m = the number of samples drawn, and the $\hat{\theta}$s are the estimated lower and upper limits of the confidence interval (Sievers, 1996). The standard error of each alpha rate is equal to $\sqrt{s(1-s)/m}$, where s is the statistical significance level and m the number of samples drawn from the simulated population (Rasmussen, 1989). Values within ±2SE (i.e., $\sqrt{.05(1-.05)/100} = .022 * 2 = .044$) may be considered within sampling error, though the discrepancy between a given error rate and the nominal value is not especially relevant to our purposes because the present concern is in locating the bootstrap method yielding the most accurate error rate per correlation type, regardless of the value's distance from the nominal rate. Nevertheless, the demarcation is useful in revealing bootstrap methods that be particularly inaccurate.

Interval ratio. Interval accuracy is determined by comparing the bootstrap confidence intervals to "true" confidence intervals. However, the distribution of m sample statistics obtained from Step 9 is not sufficient to constitute a Monte Carlo estimate of the true confidence intervals. Therefore, 10,000 samples were drawn from each derived population. This process was repeated for each sample size condition. The distribution of correlation coefficients calculated form each sample serves as an estimated "true" sampling distribution from which the quantiles $Q_{\alpha/2}$ and $Q_{1-\alpha/2}$ can be obtained. These quantiles are empirical estimates of the "true" limits of the sampling distribution of θ and were used as a standard against which to compare several indices to be described next (see King, 2000, for details).

A modification of a ratio proposed by Efron (1988) was applied to compare interval lengths. Efron's index entailed dividing the length of the parametric interval of r by a bootstrap interval (see Equation 52). As it was not the aim of the present study to compare bootstrap to parametric intervals, the length of each bootstrap interval was divided by the length of the "true" (Monte Carlo-estimated) confidence interval. While this ratio indexes the extent to which the intervals span an equal distance, it does not necessarily quantify the discrepancy between bootstrap endpoints and those of the "true" interval. Two intervals could conceivably have identical lengths but no overlap at all.

Bias. An even more useful gauge is bias. Bias quantifies the average discrepancy between the sample estimates and the parameter. If a negative value is obtained, the estimator underestimates the parameter on average. If $\text{BIAS}\left(\hat{\theta}\right) = 0$, the estimator $\hat{\theta}$ is unbiased and the sampling distribution of $\hat{\theta}$ is centered on θ. Because the present study was primarily designed to compare confidence intervals and not point estimates, the ordinary bias formula was modified such that each "true" endpoint was compared with the bootstrap-estimated endpoint:

$$\text{Interval Bias} = \frac{\sum_{i=1}^{m}\left(\left|\hat{\Xi}_{\text{lower}(i)} - \Xi_{\text{lower}}\right| + \left|\hat{\Xi}_{\text{upper}(i)} - \Xi_{\text{upper}}\right|\right)}{m}, \qquad (11)$$

where Ξ_{lower} and Ξ_{upper} are the "true" $(1-\alpha)100\%$ lower and upper confidence intervals for θ, and $\hat{\Xi}_{\text{lower}}$ and $\hat{\Xi}_{\text{upper}}$ are the bootstrap-estimated intervals. Because the parameter of interest in this

case is not the population correlation coefficient but the endpoints of the "true" interval, a new symbol is needed for referencing the theoretical endpoints. This was arbitrarily set to be Ξ (Xi) so that theta may be reserved for the usual population correlation coefficient.

Equation 11 computes the average of the absolute value of the differences between each bootstrap-estimated endpoint from the corresponding Monte Carlo-generated "true" endpoint. Such a bias indicator would seem to be particularly relevant in evaluating confidence intervals. While the distinction between upper and lower endpoint bias is obscured due to the sumation performed in the equation, such is rarely of interest.

Efficiency. The standard error (SE) of a statistic is defined as

$$SE\left(\hat{\theta} \mid x, n\right) = \hat{\sigma}_{\hat{\theta}} = \sqrt{\left(\frac{1}{m}\right)\sum_{i=1}^{m}\left[\hat{\theta}_i - \left(\frac{1}{m}\right)\sum_{i=1}^{m}\hat{\theta}_i\right]^2} \qquad (12)$$

and measures the spread of the estimates. This index is often referred to as a measure of efficiency. Equation 12 was modified similarly to that described for the bias measure.

Comparative Procedures

Correlational analysis. Although not typically reported in simulation studies, Type I error rates for the four bootstrap methods could be correlated to further explore the variable relationships. These correlations measure the consistency with which error rates for any two bootstrap methods covary across the range of simulation conditions. However, one should not infer from a large correlation that error rates for the two methods are identical. In fact, rates for one bootstrap procedure may run systematically higher or lower than another, but if the rates for each rise and fall monotonically in the same fashion across simulation conditions, then the correlation coefficient will fall near 1.0. Incidentally, this use of the Pearson r illustrates the value of the ordinary correlation for situations in which one does not wish to remove outliers.

Analysis of Variance. Although previous studies have usually compared error rates in an informal manner, a more formal statistical analysis was needed to process the large number of indices obtained. Analysis of Variance (ANOVA) was proposed as a means of quantifying the sources of variation affecting error rates. There are five systematic (nonrandom)

variance components that could influence error rate: correlation type, bootstrap method, distributional shape (g/h combination), sample size (n), and strength of population bivariate relationship (ρ).

One might be tempted to incorporate all of these factors in a five-way ANOVA as independent variables. Ignoring the interpretive difficulties inherent in such a proposal, such a partitioning is not possible because there is only one observation (error rate value) per cell of the balanced design. Instead, separate ANOVAs were computed in hopes of untangling the variable relationships. In particular, ANOVAs were computed in which (a) bootstrap method and correlation type served as predictors; (b) the distributional indicator (g/h combination) was added to the variables listed in (a); (c) strength of population correlation (ρ) was added to (a); and (d) sample size was added to (a). The drawback to such a course is that higher-order interactions cannot be assessed and quantified.

Results

1. How do confidence intervals for the robust correlations ρ_w and ρ_{pb} compare with those for ρ and its transform in terms of accuracy and stability?

Tables depicting alpha rates, bias, and interval ratios were constructed by averaging across the m samples for each combination of correlation type and bootstrap method and were broken down by distributional condition, sample size, and strength of population correlation. The data were also collapsed across the nine distributional conditions for easier viewing, but any performance attributable to distributional shape is masked by such a summary table. Tables 2-6 and Figures 1-2 display representative results. Disaggregated data and fuller explanations can be found in King (2000).

Confidence intervals formed for the robust correlations outperformed those for r and r_z across most of the specified criteria. Type I error rates generally fell closer to the nominal value, bias and efficiency were minimal, and "true" interval lengths were more faithfully reproduced by r_w and r_{pb}. Specifically, with nonnormal marginal distributions the robust rs frequently provided increased accuracy in error and bias rates. With normal distributions all four correlations functioned similarly in terms of Type I probability, though r and r_z bettered r_w and r_{pb} according to bias, at least when small samples were involved.

These findings for bias are not unanticipated because the robust calculations are intended to act as correctives when

nonnormal conditions arise. But the same trend should have emerged for Type I error rates.

The robust measures clearly surpassed r for both error probability and bias rate when contaminated and mixed distributions were examined. The divergence between robust and nonrobust methods was most noticeable under the extremely nonnormal $g = .5$, $h = .3$ distributional conditions. The differential performance was particularly evident in the bias index.

It should also be noted that neater, more consistent results were realized with the bias criterion than with Type I error rate, in general. This is probably due to the dichotomous nature of the latter (i.e., a given interval either does or does not enclose the parameter), in contrast to the ratio-scaled bias index. These dynamics may be responsible for some of the unexpected outcomes involving Type I error.

Efficiency rates varied little across correlational measures; all four evidenced similar stability. But when disparities arose, the robust correlations were more efficient.

In terms of interval length the Pearson statistics consistently underestimated the "true" (Monte Carlo simulated) endpoints, more so under nonnormal conditions. At times, such intervals were little more than half the "true" length. Interval lengths for the Winsorized correlation were a bit longer than desired, while the percentage bend correlation closely mimicked the "true" intervals in almost every instance.

2. Which bootstrap method provides the most accurate and stable confidence intervals for ρ_w? for ρ_{pb}?

No bootstrap technique emerged as unmistakably superior across a majority of conditions, though the BC and ordinary percentile methods yielded slightly more accurate intervals in some cases. The BC_a yielded results similar to the percentile and BC procedures in all but a few situations, while the adjusted bootstrap intervals were, by and large, unacceptable.

3. With small samples which bootstrap method provides the most accurate and stable confidence intervals for ρ_w? for ρ_{pb}?

Bootstrap procedures were not found to differ appreciably in terms of Type I error rate or bias for the majority of the simulation experiments. Consequently, one method cannot be selected as optimal for any specific condition, including sample size. Nonetheless, the adjusted bootstrap can be ruled out as a contender due to its relative inaccuracy, especially in terms of error probabilities. In addition, the BC_a frequently evidenced

slightly elevated bias and error rates with $\underline{r}$ and $\underline{r}_z$ and thus should probably not be applied to the ordinary correlation.

A minor trend in stability emerged at small sample sizes as interval lengths for the adjusted bootstrap ran wide, next was the BC_a, then the BC, and finally the ordinary percentile. These differences subsided with large samples in place. This trend was at least partly unexpected in that the BC and BC_a techniques should not fall systematically shorter or longer than the percentile method: The bias and variance inflation corrections may alter the intervals in either direction. But the adjusted bootstrap would be expected to form wider intervals than the ordinary bootstrap because the calculations involve a sample size correction. These dynamics account for the similarity between the adjusted and percentile ratios when large $\underline{n}$s were involved.

Theory predicts that the bootstrap modifications should diverge in accuracy for each of the three indicators (error, bias, and interval ratio). The BC_a should stand as the most accurate approach, then the BC, and, lastly, the percentile method. The sample size adjustment would likely fall between the BC and percentile techniques. But no such orderly progression was observed for the first two criteria, other than the unacceptable confidence intervals produced by the adjusted bootstrap.

4. Under extreme distributional conditions which bootstrap method provides the most accurate and stable confidence intervals for ρ_w? for ρ_{pb}?

With skewness and kurtosis elevated no systematic relationships were observed, beyond those just discussed, between distributional shape and bootstrap method with reference to accuracy or stability. This was true for both correlation coefficients.

Under contaminated and mixed distributions unstable error rates arose when BC_a intervals were combined with $\underline{r}$, but no substantial problems surfaced with the technique applied to robust correlations. Minor peaks in bias were similarly noted for the BC_a, but only when joined with $\underline{r}$ or $\underline{r}_z$. Surprisingly, the adjusted bootstrap was as accurate as the percentile and BC techniques for these atypical conditions.

Concluding Remarks

Wilcox (2001, p.46) summarized his bootstrapping results thusly, "These results paint a rather complicated picture about which method...to use." The same could be said of results from

the present study. However, three conclusions can be drawn from this fairly extensive simulation study. First, as regards the bootstrap procedures, the four methods under investigation performed comparably. The sample size adjustment to the percentile bootstrap did not help matters and may have actually inflated Type I error, bias, and efficiency rates, while the more complex BC and BC_a procedures failed to offer sizeable improvements in interval accuracy and thus are probably not worth the involved calculations.

These findings are disappointing, in one sense, because the adjustments were expected to generate greater accuracy in forming confidence intervals. On the other hand, researchers can be more confident that the ordinary percentile method is capable of delivering relatively precise confidence intervals, at least as applied to the four correlational measures under review and assuming these results are generalizable. Of course there are numerous bootstrap procedures that have recently seen development, and it is always possible that one of these may yield more accurate intervals.

One reason that the more complicated procedures did not surpass the percentile bootstrap may be due to the technical specifications of the simulation experiments. It was originally intended that 1,000 samples be drawn for each condition, but this number was reduced to 100 given excessive computational demands. While the goal here is not to fully reproduce a sampling distribution, more samples may be necessary to achieve stable asymptotic dynamics.

For the same reason, the number of bootstrap samples was reduced from 2,000 to 500. However, in this case the objective is to model the sampling distribution $F(\theta)$ by means of $\hat{F}(\hat{\theta}^*)$. Five hundred resamples are more than sufficient for estimating standard errors but too low to form tight confidence intervals. It was hoped that the sampling inaccuracy would influence all bootstrap procedures in an identical fashion. But if some methods are more dependent than others on the number of resamples necessary to form correct intervals (e.g., the BC_a), then results will be biased against the technique.

For example, a modest number of resamples may give rise to inadequate score representation in the tails of the sampling distribution. The tails are needed in developing accurate intervals, particularly when greater coverage probability is desired. A nominal alpha rate of .05 was employed throughout this experiment due to its popularity in social science research, but Efron (1988) warned against testing hypotheses using bootstrap resampling at such a strict probability level unless 1,000 or more bootstrap samples were acquired. This may be the cause of the Type I error rates not stabilizing at larger

sample sizes. Future studies should increase both *m* and *B* to larger sizes, possibly through the use of more efficient algorithms.

A second conclusion involves the similarity between the Winsorized and percentage bend correlations in interval accuracy. While neither robust correlation outperformed the other, both were shown to be more resilient than *r* (and r_z) to distributional abnormalities. In fact, the robust measures compared favorably to *r* even under the bivariate normal conditions. Wilcox (1993, 1997b) reported findings very similar to these for both the Winsorized and percentage bend correlations, respectively.

Third, Fisher's transformation did not appreciably improve either Type I error rate or bias associated with *r*, especially when averaged across all distributional conditions. It would seem that when bootstrapping the Pearson correlation, the transformation merely increases computational time without concomitantly affecting accuracy. But these results are in keeping with Seivers' (1996) conclusions about r_z.

In summary, the two robust correlations are to be preferred over the ordinary Pearson equations when resilience to outliers is needed. These new measures also compared favorably across normal distributional conditions and can be recommended for general use in cases where it is desired to obtain a measure of linear association reflecting the majority of the sample observations. None of the bootstrap methods under review were differentially accurate; each can be similarly endorsed, excepting the adjusted bootstrap. Perhaps future studies will produce interval-estimation techniques that afford heightened precision when applied to robust measures of correlation.

References

Baker, G. A. (1930). The significance of the product-moment coefficient of correlation with special reference to the character of the marginal distributions. Journal of the American Statistical Association, 25, 387-396.

Blair, R. C., & Lawson, S. B. (1982). Another look at the robustness of the product-moment correlation coefficient to population non-normality. Florida Journal of Educational Research, 24, 11-15.

Blomqvist, N. (1950). On a measure of dependence between two random variables. Annals of Mathematical Statistics, 21, 593-600.

Box, G. E. P. (1953). Non-normality and tests on variances. Biometrika, 40, 318-335.

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. American Statistician, 31, 147-50.

Bradley, J. V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 144-152.

Buckland, S. T., Garthwaite, P. H., & Lovell, H. G. (1988). [Discussion of Hall (1988)]. Annals of Statistics, 16, 963-966.

Cheriyan, K. C. (1945). Distributions of certain frequency constants in samples from non-normal populations. Sankhyâ, 7, 159-166.

Chesire, L., Oldis, E., & Pearson, E. S. (1932). Further experiments on the sampling distribution of the correlation coefficient. Journal of the American Statistical Association, 27, 121-128.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. Biometrika, 62, 531-545.

DiCiccio, T. J., & Romano, J. P. (1988). [Discussion of Hall (1988)]. Annals of Statistics, 16, 966-970.

Duncan, G. T., & Layard, M. W. (1973). A Monte-Carlo study of asymptotically robust tests for correlation. Biometrika, 60, 551-558.

Dunlap, H. F. (1931). An empirical determination of the distribution of means, standard deviations, and correlation coefficients drawn from rectangular populations. Annals of Mathematical Statistics, 2, 66-81.

Edgell, S. E., & Noon, S. M. (1984). Effects of violation of normality on the t test of the correlation coefficient. Psychological Bulletin, 95, 576-583.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. Annals of Statistics, 7, 1-26.

Efron, B. (1981b). Nonparametric standard errors and confidence intervals. Canadian Journal of Statistics, 9, 139-172.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. Biometrika, 72, 45-58.

Efron, B. (1987). Better bootstrap confidence intervals. Journal of the American Statistical Association, 82, 171-185.

Efron, B. (1988). Bootstrap confidence intervals: Good or bad? Psychological Bulletin, 104, 293-296.

Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. New York: Chapman & Hall.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika, 10, 507-521.

Fisher, R. A. (1921). On the 'probable error' of a coefficient of correlation deduced from a small sample. Metron, 1(4), 3-32.

Fisher, R. A. (1958). Statistical methods for research workers (13th ed.). Edinburgh: Oliver & Boyd.

Gayen, A. K. (1951). The frequency distribution of the product-moment correlation coefficient in random samples of any size drawn from non-normal universes. Biometrika, 38, 219-247.

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics, 28, 81-124.

Haldane, J. B. S. (1949). A note on non-normal correlation. Biometrika, 36, 467-468.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. Annals of Statistics, 16, 927-953.

Hall, P., Martin, M. A., & Schucany, W. R. (1989). Better nonparametric bootstrap confidence intervals for the correlation coefficient. Journal of Statistical Computation and Simulation, 33, 161-172.

Hampel, F. (1998). Is statistics too difficult? Canadian Journal of Statistics, 26, 497-513.

Havlicek, L. L., & Peterson, N. L. (1976). Robustness of the Pearson correlation against violations of assumptions. Perceptual and Motor Skills, 43, 1319-1334.

Havlicek, L. L., & Peterson, N. L. (1977). Effect of the violation of assumptions upon significance levels of the Pearson r. Psychological Bulletin, 84, 373-377.

Hey, G. B. (1938). A new method of experimental sampling illustrated on certain non-normal populations. Biometrika, 30, 68-80.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distributions. In D. C. Hoaglin, F. Mosteller, & J. Tukey (Eds.) Exploring data tables, trends, and shapes (pp. 461-513). New York: Wiley.

Huber, P. J. (1972). Robust statistics: A review. Annals of Mathematical Statistics, 43, 1041-1067.

Huber, P. J. (1981). Robust Statistics. New York: Wiley.

King, J. E. (2000). Bootstrapping robust measures of association (Doctoral dissertation, Texas A&M University, 2000). Dissertation Abstracts International, 61(11), 5948B.

Kowalski, C. J. (1972) On the effects of nonnormality on the distribution of the sample product-moment correlation coefficient. Applied Statistics, 21, 1-12.

Lunneborg, C. E., (1985). Estimating the correlation coefficient: The bootstrap approach. Psychological Bulletin, 98, 209-215.

Lunneborg, C. E., (2000). Data analysis by resampling: Concepts and applications. Pacific Grove, CA: Brooks/Cole.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105, 156-166.

Mooney, C. Z., & Duval, R. D. (1993). Bootstrapping: A nonparametric approach to statistical inference. Newbury Park, CA: Sage.

Nair, A. N. K. (1941). Distribution of Student's "t" and the correlation coefficient in samples from non-normal populations. Sankhyâ, 5, 383-400.

Norris, R. C., & Hjelm, H. F. (1961). Nonnormality and product-moment correlation. Journal of Experimental Education, 29, 261-270.

Pearson, E. S. (1929). Some notes on sampling tests of two variables. Biometrika, 21, 337-360.

Pearson, E. S. (1931). The test of significance for the correlation coefficient. Journal of the American Statistical Association, 26, 128-134.

Pearson, E. S. (1932). The test of significance for the correlation coefficient: Some further results. Journal of the American Statistical Association, 27, 424-426.

Rasmussen, J. L. (1986). An evaluation of parametric and non-parametric tests on modified and non-modified data. British Journal of Mathematical and Statistical Psychology, 39, 213-220.

Rasmussen, J. L. (1987). Estimating correlation coefficients: Bootstrap and parametric approaches. Psychological Bulletin, 101, 136-139.

Rasmussen, J. L. (1988). "Bootstrap confidence intervals: Good or bad": Comments on Efron (1988) and Strube (1988) and further evaluation. Psychological Bulletin, 104, 297-299.

Rasmussen, J. L. (1989). Computer-intensive correlational analysis: Bootstrap and approximate randomization techniques. British Journal of Mathematical and Statistical Psychology, 42, 103-111.

Rider, P. P. (1932). On the distribution of the correlation coefficient in small samples. Biometrika, 24, 382-402.

Sheppard, W. F. (1899). On the application of the theory of error to cases of normal distribution and normal correlation. Philosophical Transactions of the Royal Society of London, 192(Pt. A), 101-167.

Sievers, W. (1996). Standard and bootstrap confidence intervals for the correlation coefficient. British Journal of Mathematical and Statistical Psychology, 49, 381-396.

Srivastava, M. S., & Awan, H. M. (1984). On the robustness of the correlation coefficient in sampling from a mixture of two bivariate normals. Communications in Statistics-Theory and Methods, 13, 371-382.

Staudte, R. G., & Sheather, S. J. (1990). Robust estimation and testing. New York: Wiley.

Stigler, S. M. (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation 1885-1920. Journal of the American Statistical Association, 68, 872-879.

Strube, M. J. (1988). Bootstrap Type I error rates for the correlation coefficient: An examination of alternate procedures. Psychological Bulletin, 104, 290-292.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61(4), 361-377.

Thompson, B., & Snyder, P. A. (1997). Statistical significance testing practices in the Journal of Experimental Education. Journal of Experimental Education, 66, 75-83.

Thompson, B. (1999). Improving research clarity and usefulness with effect size indices as supplements to statistical significance tests. Exceptional Children, 65, 329-337.

Thompson (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. Journal of Experimental Education, 70, 80-93.

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin et al. (Eds.), Contributions to probability and statistics. Stanford, CA: Stanford University Press.

Vacha-Haase, T., Nilsson, J.E., Reetz, D.R., Lance, T.S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. Theory & Psychology, 10, 413-425.

Wilcox, R. R. (1990). Comparing variances and means when distributions have non-identical shapes. Communications in Statistics-Simulation and Computation, 19, 155-173.

Wilcox, R. R. (1991). Bootstrap inferences about the correlation and variances of paired data. British Journal of Mathematical and Statistical Psychology, 44, 379-382.

Wilcox, R. R. (1993). Some results on a Winsorized correlation coefficient. British Journal of Mathematical and Statistical Psychology, 46, 339-349.

Wilcox, R. R. (1994). The percentage bend correlation coefficient. Psychometrika, 59, 601-616.

Wilcox, R. R. (1996). A review of some recent developments in robust regression. British Journal of Mathematical and Statistical Psychology, 49, 253-274.

Wilcox, R. R. (1997a). An introduction to robust estimation and hypothesis testing. San Diego, CA: Academic Press.

Wilcox, R. R. (1997b). Tests of independence and zero correlations among p random variables. Biometrical Journal, 39, 183-193.

Wilcox, R. R. (1998a). How many modern discoveries have been lost by ignoring modern statistical methods? American Psychologist, 53, 300-314.

Wilcox, R. R. (1998b). [Reply]. British Journal of Mathematical and Statistical Psychology, 51, 55-62.

Wilcox, R. R., & Muska, J. (2001). Inferences about correlations when there is heteroscedasticity. British Journal of Mathematical & Statistical Psychology, 54, 39-47.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604.

Zeller, R. A., & Levine, Z. H. (1974). The effects of violating the normality assumption underlying r. Sociological Methods and Research, 2, 511-519.

Footnotes

[1]These terms are not uniformly used in the literature.

[2]Actually, it is assumed that the difference distributions, $F(\hat{\theta} - \theta)$ and $\hat{F}(\hat{\theta}^* - \hat{\theta})$, are equivalent.

[3]Not described in the present paper.

Table 1

Illustrative Parameter Values for the *g* and *h* Distribution

g	h	M	SD	Skewness	Kurtosis
0	0	0.000	1.000	0.001	0.000
0	.1	0.000	1.182	0.004	2.572
0	.3	0.000	1.966	1.082	218.827
.2	0	0.102	1.030	0.613	0.666
.2	.1	0.119	1.227	1.049	4.871
.2	.3	0.171	2.152	11.955	1,151.082
.5	0	0.268	1.211	1.758	5.918
.5	.1	0.313	1.505	3.391	38.287
.5	.3	0.465	3.035	21.175	1,413.162

Note. Entries based on 1,000,000 simulated observations.

Skewness and kurtosis were computed as: skewness = $\underline{m}_3 / \underline{m}_2(\underline{m}_2)^{1/2}$, kurtosis = $(\underline{m}_4 / \underline{m}_2^2) - 3$, where $\underline{m}_a = \sum_{i=1}^{N} (x_i - \mu)^a$.

Table 2

Type I Error Rates Averaged Across All g and h Distributional Conditions

	n = 20				*n* = 50				*n* = 100				*n* = 250			
Method	r	r_z	r_w	r_{pb}	r	r_z	r_w	r_{pb}	r	r_z	r_w	r_{pb}	r	r_z	r_w	r_{pb}
ρ = 0																
Percentile	.06	.06	.03	.04	.07	.07	.06	.07	.05	.05	.05	.05	.07	.07	.04	.04
Adjusted	.06	.06	.03	.04	.07	.07	.06	.07	.05	.05	.05	.05	.07	.07	.04	.04
BC	.05	.05	.03	.05	.07	.07	.06	.06	.05	.05	.06	.05	.06	.06	.03	.04
BCA	.05	.04	.03	.05	.07	.07	.05	.06	.06	.06	.06	.05	.07	.07	.04	.04
ρ = .4																
Percentile	.11	.11	.03	.07	.08	.08	.04	.06	.07	.07	.04	.04	.08	.08	.05	.05
Adjusted	.11	.11	.04	.08	.08	.08	.04	.06	.08	.08	.04	.04	.08	.08	.05	.05
BC	.08	.08	.03	.06	.08	.08	.03	.06	.08	.08	.04	.04	.09	.09	.05	.05
BCA	.09	.08	.04	.07	.09	.09	.04	.06	.10	.10	.04	.03	.11	.11	.04	.05
ρ = .8																
Percentile	.09	.09	.06	.07	.06	.06	.06	.06	.06	.06	.06	.04	.07	.07	.05	.05
Adjusted	.15	.15	.09	.12	.06	.06	.06	.06	.08	.08	.07	.07	.08	.08	.04	.05
BC	.10	.10	.05	.05	.06	.06	.06	.07	.07	.07	.06	.04	.08	.08	.06	.05
BCA	.12	.12	.06	.07	.07	.07	.05	.06	.10	.10	.06	.04	.09	.09	.06	.06

Note. Underscored values are greater than two standard errors beyond the nominal .05 level.

Table 3

Analysis of Variance for Type I Error Rate by Correlation Type and Bootstrap Method

Source	df	F	p	η^2
Model	15	11.028	<.001	.088
CORR	3	50.511	<.001	.081
BOOT	3	2.735	.042	.004
CORR * BOOT	9	.631	.772	.003
Error	1712	(.002)		
Total	1727			

Note. Value enclosed in parentheses represents mean square error.

Table 4

Interval Bias Averaged Across All g and h Distributional Conditions

		n = 20				n = 50				n = 100				n = 250			
	Method	r	r_z	r_w	r_{pb}	r	r_z	r_w	r_{pb}	r	r_z	r_w	r_{pb}	r	r_z	r_w	r_{pb}
ρ = 0																	
	Percentile	.33	.33	.30	.31	.24	.24	.22	.22	.17	.17	.16	.16	.11	.11	.10	.10
	Adjusted	.36	.36	.35	.34	.24	.24	.23	.23	.17	.17	.16	.16	.11	.11	.10	.10
	BC	.32	.32	.31	.31	.23	.23	.22	.22	.16	.16	.16	.16	.11	.11	.10	.10
	BCA	.34	.33	.31	.31	.24	.24	.22	.22	.17	.17	.16	.16	.11	.11	.10	.10
ρ = .4																	
	Percentile	.36	.36	.33	.33	.24	.24	.20	.20	.21	.21	.15	.15	.15	.15	.10	.09
	Adjusted	.38	.38	.37	.37	.24	.24	.21	.21	.21	.21	.15	.15	.15	.15	.10	.09
	BC	.36	.36	.33	.32	.24	.24	.21	.20	.21	.21	.15	.14	.16	.16	.10	.10
	BCA	.38	.37	.33	.33	.26	.25	.21	.20	.23	.23	.15	.15	.17	.17	.10	.10
ρ = .8																	
	Percentile	.26	.26	.25	.23	.17	.17	.14	.13	.13	.13	.09	.08	.10	.10	.06	.05
	Adjusted	.31	.31	.31	.30	.17	.17	.15	.15	.13	.13	.09	.09	.10	.10	.06	.06
	BC	.26	.26	.28	.24	.17	.17	.14	.14	.14	.14	.09	.09	.11	.11	.06	.06
	BCA	.28	.28	.28	.25	.18	.18	.15	.15	.15	.15	.09	.09	.12	.12	.06	.06

Table 5

Analysis of Variance for Bias by Correlation Type and Bootstrap Method

Source	df	F	p	η^2
Model	15	3.497	<.001	.030
CORR	3	15.558	<.001	.026
BOOT	3	1.551	.199	.003
CORR * BOOT	9	.125	.999	.001
Error	1712	(.010)		
Total	1727			

Note. Value enclosed in parentheses represents mean square error.

Table 6

Confidence Interval Ratios Averaged Across All g and h Distributional Conditions

	n = 20				n = 50				n = 100				n = 250			
Method	r	r_z	r_w	r_{pb}	r	r_z	r_w	r_{pb}	r	r_z	r_w	r_{pb}	r	r_z	r_w	r_{pb}
ρ = 0																
Percentile	.91	.91	1.11	1.02	.91	.91	1.04	1.00	.92	.92	1.03	1.00	.93	.93	1.01	.99
Adjusted	.98	.98	1.20	1.10	.94	.94	1.07	1.03	.94	.94	1.04	1.02	.94	.94	1.01	1.00
BC	.92	.92	1.11	1.02	.91	.91	1.04	1.00	.92	.92	1.03	1.00	.93	.93	1.01	.99
BCA	.92	.92	1.11	1.02	.92	.92	1.04	1.00	.93	.93	1.02	1.00	.94	.94	1.01	.99
ρ = .4																
Percentile	.84	.84	1.09	1.00	.84	.84	1.04	.99	.92	.92	1.03	1.00	.86	.86	1.01	1.00
Adjusted	.91	.91	1.17	1.08	.86	.86	1.07	1.02	.93	.93	1.04	1.02	.87	.87	1.02	1.00
BC	.86	.86	1.11	1.02	.84	.84	1.05	1.00	.91	.91	1.02	1.00	.86	.86	1.01	1.00
BCA	.85	.86	1.11	1.02	.84	.84	1.05	1.01	.92	.92	1.03	1.00	.86	.86	1.01	1.00
ρ = .8																
Percentile	.81	.81	1.08	.98	.85	.85	1.05	1.02	.81	.81	1.00	.98	.84	.84	1.01	1.00
Adjusted	.87	.87	1.16	1.06	.88	.88	1.08	1.05	.83	.83	1.01	.99	.84	.84	1.01	1.00
BC	.85	.85	1.17	1.04	.87	.87	1.08	1.04	.82	.82	1.01	.99	.84	.84	1.01	1.00
BCA	.83	.86	1.16	1.05	.85	.87	1.10	1.06	.81	.82	1.02	1.01	.84	.85	1.01	1.01

Figure 1. Mean Type I error rate by correlation type and bootstrap method. Reference line indicates the nominal alpha rate of .05.

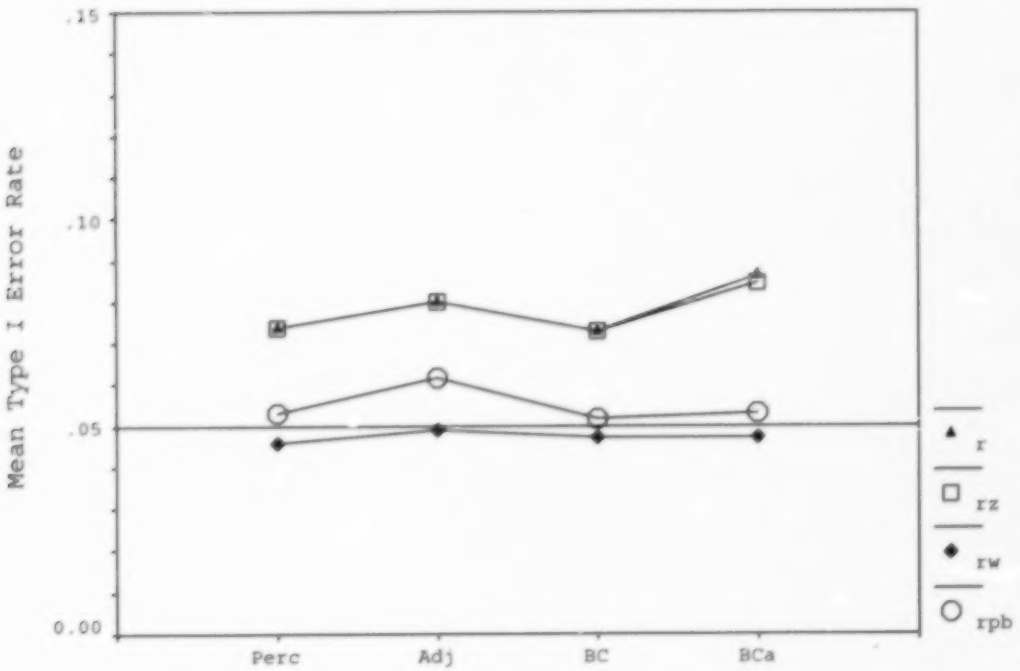

Figure 2. Mean bias by correlation type and bootstrap method.